Evidence Synthesis for Decision Making 7: A Reviewer's Checklist

A. E. Ades, PhD, Deborah M. Caldwell, PhD, Stefanie Reken, MSc, Nicky J. Welton, PhD, Alex J. Sutton, PhD, Sofia Dias, PhD

This checklist is for the review of evidence syntheses for treatment efficacy used in decision making based on either efficacy or cost-effectiveness. It is intended to be used for pairwise meta-analysis, indirect comparisons, and network meta-analysis, without distinction. It does not generate a quality rating and is not prescriptive. Instead, it focuses on a series of questions aimed at revealing the assumptions that the authors of the synthesis are expecting readers to accept, the adequacy of the arguments authors advance in support of their position, and the need for further analyses or sensitivity analyses. The checklist is intended primarily for those who review evidence syntheses, including indirect comparisons and network meta-analyses, in the context of decision making but will also be of value to those submitting syntheses for review, whether to decision-making bodies or journals. The checklist has 4 main headings: A) definition of the decision problem, B) methods of analysis and presentation of results, C) issues specific to network synthesis, and D) embedding the synthesis in a probabilistic cost-effectiveness model. The headings and implicit advice follow directly from the other tutorials in this series. A simple table is provided that could serve as a pro forma checklist. ***Key words:*** *cost-effectiveness analysis; Bayesian meta-analysis; multiparameter evidence synthesis; meta-analysis.* ***(Med Decis Making 2013;33:679–691)***

This tutorial article sets out a practical checklist intended primarily for those who review evidence syntheses, including indirect comparisons and network meta-analyses (NMAs), in the context of decision making. The checklist can also be used by those preparing such evidence syntheses. It consists of a set of systematic criteria by which an independent reviewer can assess whether the synthesis meets the requirements elaborated in the other tutorials in this series.[1–6]

Our assumption is that the purpose of the synthesis is to obtain a comparison, for purposes of efficacy and/or cost-effectiveness, of a *prespecified set of treatments* in patients with *a prespecified set of characteristics*. The purpose of this restriction is to tie the checklist firmly to the ''decision-making'' context in which a clinician or the policy maker has a particular set of patients and precisely defined treatments in mind. This is the level at which reimbursement authorities typically operate and in which clinicians are interested. However, not all evidence synthesis is conceived in precisely this way. A number of systematic reviews and meta-analyses are carried out with a primary objective of summarizing literature on a particular treatment comparison or set of comparisons, often in a broader range of patient groups. The proposed checklist is not primarily intended for that broader form of review, although it may be highly relevant to it.

We make no attempt to produce a summary quality rating of the synthesis. A numerical or qualitative rating does not provide the information that decision makers require to determine whether a submitted synthesis represents an adequate basis for the decision they are charged with making. Similarly, a numerical or quality rating does not help an editorial board decide whether to accept a paper based on an evidence synthesis. Instead, we see the checklist as a way of assessing whether the synthesis and

Received 16 October 2012 from the School of Social and Community Medicine, University of Bristol, Bristol, UK (AEA, DMC, NJW, SD); National Institute for Health and Clinical Excellence, London, UK (SR); and Department of Health Sciences, University of Leicester, Leicester, UK (AJS). This series of tutorial papers were based on Technical Support Documents in Evidence Synthesis (available from http://www.nicedsu.org.uk), which were prepared with funding from the NICE Decision Support Unit. The views, and any errors or omissions, expressed in this document are of the authors only. Revision accepted for publication 20 December 2012.

Address correspondence to Sofia Dias, School of Social and Community Medicine, University of Bristol, Canynge Hall, 39 Whatley Road, Bristol BS8 2PS, UK; e-mail: s.dias@bristol.ac.uk.

DOI: 10.1177/0272989X13485156

the conclusions drawn from it are a fair reflection of what can be concluded from the existing evidence, whatever its quality. The completed checklist would form the basis of a reviewer report to the decision maker or journal editor. However, the checklist also tells those submitting a synthesis precisely what are the critical issues they will be expected to clarify, what arguments and evidence they may be called upon to marshal, and the sensitivity analyses they may be asked to undertake.

Our objective is to provide a framework for open discussion of whether a convincing argument has been made, albeit based on data that may be limited and imperfect. Similarly, there is no attempt to quantify the ''strength of evidence'' or suggest a ''strength of recommendation.''[7,8] A convincing argument can be developed from poor evidence, and the strength of evidence should be fully reflected in the credible interval attached to it, which should incorporate not only sampling error but uncertainty due to bias adjustment or due to the uncertain relevance of the available data.[3] If decisions are based on cost-effectiveness, the strength of recommendation is better expressed through the commonly used metrics such as incremental cost-effectiveness ratios, cost-effectiveness acceptability curves, and probability that a strategy is optimal, given the model and a threshold willingness to pay.[9]

RELATION TO OTHER CHECKLISTS

Throughout this tutorial series, we have defined NMA as an extension of pairwise meta-analysis.[2] A key assumption is that for any pair of treatments under consideration, the true relative treatment effects are either identical (fixed effect model) or exchangeable (random effect [RE] model), across *all* the trials in the set. This identity or exchangeability requirement is present for any pair of treatments X and Y. It is therefore not strictly correct to claim that NMA requires *extra* assumptions of ''trial similarity'' and ''consistency,'' *additional* to assumptions that are required in pairwise meta-analysis, as has been occasionally claimed.[10–12] But this is not to say that these properties are unimportant. On the contrary, the fact that pairwise and network meta-analysis are so close in their underlying assumptions only serves to emphasize that all the ''good-practice'' advice that is incorporated in existing guidance[13,14] and checklists[15–17] available for pairwise meta-analysis is also the essential guarantor of adequacy in NMA. Equally, it highlights that these assumptions deserve scrutiny in the context of pairwise synthesis, particularly as there is even less possibility of checking them within the data at hand.

Rather than duplicate existing guidelines for conducting or reporting systematic reviews,[13,17,18] we assume that these have been followed. Most items in the proposed checklist apply to both pairwise and network meta-analyses. The only issues that come exclusively under the heading of network synthesis are connectedness of networks, inconsistency, and software implementation. Setting these aside, our checklist tends to be *more* restrictive than existing guidelines in handling effect modifiers and potential effect modifiers, as this is likely to be inherent in the decision question. In other respects, our approach is less restrictive, in that we would encourage syntheses of multiple outcomes within a single coherent model[19–22] rather than a separate synthesis for each outcome.

Although this checklist covers similar items to that of the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) taskforce,[23,24] it has been designed to be more suited to inform an actual decision-making process rather than guide academic paper submissions.

HOW TO INTERPRET AND USE THE CHECKLIST

Our objective in providing a checklist is to provide guidance on what questions should be asked of an evidence synthesis by a reviewer or any other reader. The suggested checklist expresses a record of ''fact'' about a synthesis, as well as its conduct and assumptions, but also provides room for comments. These may include expressions of doubt about assumptions or interpretations of evidence and may point to the need for further analyses or sensitivity analyses.

In certain cases, relatively strong assumptions may be necessary due to the lack of evidence. Furthermore, empirical approaches to testing those assumptions may be limited by the data available. A thorough and transparent discussion of all assumptions and their implications should be provided. The checklist allows the reviewer to comment on whether the assumptions are reasonable and adequately justified and to indicate whether the issue in question has been adequately addressed. For example, in reply to a question on whether additional modeling assumptions were made, the reviewer may answer with a ''tick'' adding comments such as ''no additional assumptions'' or ''additional assumptions justified'' or put a ''cross'' with a comment indicating that the assumptions are questionable.

The checklist comes in 4 sections. It begins with a set of considerations relating to the definition of the decision problem, comparators, and target patient population or populations, but also what is already known about the potential role of known or unknown effect modifiers. The second section turns to the data analysis methods and to the results. The third section examines issues specific to NMA: connectedness and inconsistency. A final section touches on uncertainty propagation in the cost-effectiveness analysis (CEA). We refer to the other tutorials in this series throughout for further details. A downloadable version of the checklist is available from www.nicedsu.org.uk.

The checklist has been tested on published network meta-analyses—the checklist was easy to use, and problems with the methodology and/or the patient population used in the reports were successfully identified.

THE CHECKLIST

A. Definition of the Decision Problem

A1. Target Population for Decision

A1.1. Has the target patient population for decision been clearly defined?

Reviewers should note whether the target population is clearly defined and whether there is more than 1 population and, therefore, more than 1 decision involved. Each decision would require its own CEA.

A2. Comparators

A2.1. Decision comparator set: Have all the appropriate treatments in the decision been identified?

The *decision comparator set* of treatments includes all the treatments to be compared, as identified in the scoping exercise.[1,25] Ideally, this should include all the candidate treatments for the target population in question.

A2.2. Synthesis comparator set: Are there additional treatments in the synthesis comparator set that are not in the decision comparator set? If so, is this adequately justified?

The *synthesis comparator set* consists of all the treatments in the decision set plus any other treatments used in the synthesis.[1,25] One reason for adding treatments to the synthesis set might be to make a connected network.[1,26] It is sometimes possible to extend the comparison set still further,[27] although this should not be regarded as the base-case analysis. The advantages of this extension are the increased potential to check consistency, the potential to reduce uncertainty by including more evidence, and the fact that the final results will be more robust and less sensitive to the inclusion of any individual trial. The potential disadvantage is increased risk of heterogeneity in patient populations. If expansion of the network leads to increased heterogeneity, this may result in *increased* uncertainty in estimates from RE models.[28] The increased uncertainty may be an appropriate reflection of the true state of affairs, and the increased robustness conferred by a larger ensemble of data may be seen as outweighing this. Another reason for extending the set of comparators is to be able to include trials that provide additional information on the relationship between outcomes.[5]

A3. Trial Inclusion/Exclusion

A3.1. Is the search strategy technically adequate and appropriately reported?

To minimize bias in the systematic review, a thorough search of the literature should be conducted. This should be reported in sufficient detail so that it can be judged and reproduced, if required.[13] Methods for review protocols and reporting should be adopted according to current best practice.[13,17,18]

A3.2. Have all trials involving at least 2 of the treatments in the synthesis comparator set been included?

If some have been excluded, which are they, and have adequate reasons been given? Should sensitivity to inclusion/exclusion of these studies individually and/or together be provided?

There is no specific reason to rule out trials on the basis of their size or design, for example, because they were noninferiority trials. All things being equal, these design features should have no impact on the validity of the estimates obtained, only their variance.[25] Possibly, a case could be made for ruling out smaller trials if there was reason to suspect publication bias or small-study bias, but this should be based on a formal analysis, with examination of funnel plots or other methods.[3,29,30] Crossover or cluster-randomized trials should also be included if they have been analyzed and reported appropriately.

Trials that were stopped early (under a protocol with prespecified early stopping rules) should also be included, without adjustment for early stopping.[31,32] Multiarm trials involving at least 2 of the treatments in the synthesis comparator set should also be included. Treatments outside the synthesis

comparator set can be excluded, as they contribute nothing to the analysis. Single-arm studies cannot be included in a relative efficacy analysis.

A3.3. Have all trials reporting relevant outcomes been included?

If different trials report different but clearly related outcomes or the same outcome has been reported in different ways (e.g., as hazard ratios or median time to an event) or at different time points, a synthesis incorporating different reporting formats or test instruments within a single coherent model should be undertaken. Methods for combining data reported in different formats, such as shared parameter models,[2] should be considered.[19,21]

A3.4. Have additional trials been included? If so, is this adequately justified?

Trials that would not fall within the strict target definition of patients or treatments may be included, if the trial population, treatment protocol, or dosing is ''similar'' to those within the decision problem. The key assumption, that the relative treatment effects are identical or exchangeable with those in the target population, must be explicitly addressed,[3] and sensitivity analyses excluding these studies should be considered. If further trials have been included, it needs to be established that there has been no arbitrary selection from among a set of eligible trials.

A4. Treatment Definition

A4.1. Are all the treatment options restricted to specific doses and co-treatments, or have different doses and co-treatments been ''lumped'' together? If the latter, is it adequately justified?

In a decision-making context, the doses and treatment regimes being considered for every treatment in the decision set are almost always tightly defined.[1,26] The practice of ''lumping''[33] doses or co-treatments together generally makes no sense in decision making, unless the variations in dose or co-treatment are so small that clinicians would agree that the variation has no material impact on efficacy.[4] Lumping over different doses or co-treatments introduces heterogeneity and inconsistency.[34–38] If different doses or different co-treatments are considered to have the same efficacy, this should be explicitly addressed and justified.[3]

A4.2. Are there any additional modeling assumptions?

It is open to investigators to fit, for example, dose-response models[39] or to fit models in which the effect of a complex intervention can be derived from the effects of the subcomponents.[40] Evidence in the literature that bears on the validity of such models in the current context should be reviewed and their a priori clinical or scientific plausibility discussed. Evidence in the form of goodness of fit of alternative models should be presented.[2]

A5. Trial Outcomes and Scale of Measurement Chosen for the Synthesis

A5.1. Where alternative outcomes are available, has the choice of outcome measure used in the synthesis been justified?

Several different outcomes may be reported in a set of trials and at more than 1 follow-up time. If a single outcome or follow-up time is selected, this should be justified. A coherent synthesis of several outcomes should give more robust results (e.g., probit or logit models for ordered categorical outcomes[2]), but the validity of such models should be established by citing previous literature and/or by examining their validity and goodness of fit.[2]

A5.2. Have the assumptions behind the choice of scale been justified?

The choice of outcome measure that forms the basis for the synthesis (e.g., log odds ratio, log relative risk, log hazard ratio, risk difference) should be justified, as there is a strong assumption that the true effects are linear on the chosen scale.[2] Analysis of rate outcomes in most cases assumes constant hazards over time in each trial arm and a proportional hazards treatment effect. The plausibility of constant hazards, particularly when trial follow-up times vary greatly, needs to be discussed. Conversely, the use of logit models for probability outcomes in studies with different follow-up times implies very different assumptions. One option is to assume that all outcome events that are going to occur will have occurred before the observation period in the trial has ended, regardless of variation between studies in follow-up time. Another is to assume a proportional odds model, which implies a complex form for the hazard rates.[41] The clinical plausibility of these assumptions should be discussed and supported either by citing relevant literature or by examination of evidence on changes in outcome rate over the period of follow-up.

A6. Patient Population: Trials with Patients outside the Target Population

A6.1. Do some trials include patients outside the target population? If so, is this adequately justified?

A6.2. What assumptions are made about the impact or lack of impact this may have on the relative treatment effects? Are they adequately justified?

A6.3. Has an adjustment been made to account for these differences? If so, comment on the adequacy of the evidence presented in support of this adjustment and on the need for a sensitivity analysis.

Some trials have a patient population that differs somewhat from the target population for decision. If these trials are included, investigators must be explicit about what they are assuming and give a reasoned argument justifying their approach. One alternative is to say that the patients may have different characteristics, but that would not be expected to affect the treatment effects. The other is to include some form of *adjustment* in the analysis, to obtain an adjusted estimate that would represent the treatment effect expected in the target population. This adjustment could be based on data from another trial or cohort study, expert elicitation,[42] meta-regression,[3] or a bias adjustment model.[3,42–45]

A7. Patient Population: Heterogeneity within the Target Population

A7.1. Have potential modifiers of treatment effect been considered?

This may be based on clinical opinion or on a separate review of the literature.

A7.2. Are there apparent or potential differences between trials in their patient populations, albeit within the target population? If so, has this been adequately taken into account?

Although the patient population of every trial appears to lie *within* the definition of the target population, there may still be heterogeneity between the trial populations—perhaps based on age, referral pattern, previous treatment, or disease severity. One option for the investigator is to consider that neither the relative treatment nor the baseline treatment effects are influenced by the patient heterogeneity. A second option is that the relative effects remain unchanged, but baseline effects are different. This would lead to a form of subgroup analysis on baselines and potentially to different decisions being taken for different patient groups (see section B3). A final possibility would be that the relative effects vary. This would lead, potentially, to a subgroup analysis based on a covariate that modified the treatment effect.[3] This would require discussion of any a priori clinical rationale for a subgroup effect and empirical evidence for it in the literature.

A8. Risk of Bias

A8.1. Is there a discussion of the biases to which these trials, or this ensemble of trials, are vulnerable?

A8.2. If a bias risk was identified, was any adjustment made to the analysis and was this adequately justified?

An account should be given of the characteristics of each of the individual trials that could be associated with bias and also the possibility of publication or small-study biases attaching to the ensemble of trials. There should also be an account of the potential impact trial quality could have on the synthesis results.[46] Biases associated with indicators of trial quality are a particular concern, as these may act to increase treatment effect.[47–52] Methods for adjusting for these biases should be considered.[3]

A9. Presentation of the Data

A9.1. Is there a clear table or diagram showing which data have been included in the base-case analysis?

A network diagram is a useful way of showing the structure of the evidence. The actual data used in the base-case analysis (trial first author and date, outcomes, treatments compared, and covariates if relevant) should be set out in a table. Good practice examples are given in other tutorials in this series and their appendices.[1,2,4]

A9.2. Is there a clear table or diagram showing which data have been excluded and why?

Details of all trials and outcomes not considered for the analysis should be detailed in a table or diagram, along with reasons.[17] In the interest of transparency, a note should be made of other potentially relevant data available, such as information on related outcomes, outcomes reported at more than one time point, or survival curves.

B. Methods of Analysis and Presentation of Results

B1. Meta-Analytic Methods

B1.1. Is the statistical model clearly described?

Reviewers should be provided with a precise description of the meta-analytic method used. The model should either be presented in algebraic form, or a citation should be provided to the statistical model being assumed. If a Bayesian analysis is used, details on priors, convergence, and number of iterations should also be given.[14]

Reviewers should check that the meta-analysis method used is statistically sound for the data set at hand. For example, the addition of 0.5 to zero cell counts can materially bias the estimated treatment effects. If the treatment effects are strong and the event is common or there is large sample size imbalance between the groups, the Peto method should be avoided.[13] Fixed effect estimators should not be used without considering possible heterogeneity. Further guidance is provided in standard texts on meta-analysis.[53–55]

B1.2. Has the software implementation been documented?

The name of the software module and package used for statistical analysis should be given, and any additional computer code should be provided, to ensure that analyses can be replicated. If confidentiality issues exist, fictitious data can be used.

B2. Heterogeneity in the Relative Treatment Effects

B2.1. Have numerical estimates been provided of the degree of heterogeneity in the relative treatment effects?

An assessment should be made of the degree of heterogeneity in relative treatment effects for each set of pairwise comparisons. Tests of the null hypothesis of homogeneity,[13] the I^2 statistic,[56] or estimates of the between-trial variation in an RE model are useful. The latter are particularly valuable as they can be compared with the estimated treatment effects.[3]

B2.2. Has a justification been given for choice of random or fixed effect models? Should sensitivity analyses be considered?

The results of such analyses can be used, in part, to justify the choice of RE models. In a Bayesian context, deviance information criterion statistics can also be used for this.[2]

B2.3. Has there been an adequate response to heterogeneity?

If there is substantial heterogeneity in relative treatment effects, the role of known or unknown covariates and potential for random biases, as well as the possible role of bias adjustment or control for variation by covariate adjustment, should be discussed.[3] Covariate adjustment will usually have implications for the decision question as it raises the possibility of different treatment effects in different patient groups.

B2.4. Does the extent of unexplained variation in relative treatment effects threaten the robustness of conclusions?

As the between-studies standard deviation approaches the average treatment effect in magnitude, it is legitimate to ask how this affects the validity of conclusions. One might be confident that the mean treatment effect in an RE model is greater than zero while still being quite uncertain about whether the treatment effect will be positive in a future instance.[3] To interpret such heterogeneity in a decision context, one suggestion is that the predictive distribution of the treatment effect in a new trial is the appropriate input in a decision analysis, rather than the mean effect.[3,57–59] This could be considered to better represent the uncertainty in the treatment effect, without materially changing the expected treatment effect.

B2.5. Has the statistical heterogeneity between baseline arms been discussed?

The extent of heterogeneity in the baseline arms should be discussed, as it may provide information on the heterogeneity of the patient populations. Heterogeneity in baselines should lead to reexamination of trial inclusion criteria and the risk of heterogeneous treatment effects.

B3. Baseline Model for Trial Outcomes

B3.1. Are baseline effects and relative effects estimated in the same model? If so, has this been justified?

In this tutorial series,[5] we have strongly recommended that the model for the relative treatment effects is independent of the model for the baseline model. The intention is to avoid biasing the relative effect model by choosing a baseline model whose assumptions are not correct. The Bayesian approach presented[2] is based on the likelihoods of the trial arms rather than likelihoods of relative effects. Vague unrelated priors are assumed for the ''baseline'' arm of each trial, and the relative effects are modeled. Simultaneous modeling of baseline and relative effects should generally be avoided unless a clear reason can be given.[5]

B3.2. Has the choice of studies used to inform the baseline model been explained?

The source of data used for the baseline model should be explained and justified.[5] Use of the placebo arms from the available studies or from a suitable subset of the included studies are 2 options, but external data could also be considered. The source, or sources, of data that best represent the outcome that would be obtained with the standard treatment in the target population should be used. If several sources of data are available, methods for averaging them should be justified. Where

heterogeneous data are used, the use of the predictive distribution should be considered.[5]

B4. Presentation of Results of Analyses of Trial Data

B4.1. Are the relative treatment effects (relative to a placebo or ''standard'' comparator) tabulated, alongside measures of between-study heterogeneity if an RE model is used?
B4.2. Are the absolute effects on each treatment, as they are used in the CEA, reported?

Guidance on what results should be presented is available in the tutorials in this series.[1–3] A table with the results based only on direct evidence and on the full network analysis is very informative,[38] as are other graphical and tabular displays[60,61] such as rank-o-grams,[22,62] which plot the probabilities that each treatment is the best, second best, and so on.

B5. Synthesis in Other Parts of the Natural History Model

The relative treatment effect model and the baseline model are both based on the short-term outcomes that are reported in trials. However, in most CEA models, there is a need to project this ''downstream'' so that the natural history reflects posttrial outcomes.

B5.1. Is the choice of data sources to inform the other parameters in the natural history model adequately described and justified?
B5.2. In the natural history model, can the longer-term differences between treatments be explained by their differences on randomized trial outcomes?

Construction and interpretation of natural history models are greatly facilitated when the values of parameters ''downstream'' from the trial outcomes are independent of treatment. When these parameters do depend on treatment, they will often be informed from observational evidence. The use of observational evidence to drive differences in relative treatment effects needs to be carefully justified and explained. Potential sources of bias should be discussed.

C. Issues Specific to Network Synthesis

The need for a detailed description of the methods and software implementation applies equally to indirect comparisons and NMA.

C1. Adequacy of Information on Model Specification and Software Implementation

For NMA and indirect comparisons, the WinBUGS code for Bayesian evidence synthesis set out in this series[2] is a recommended option. The STATA package mvmeta[63] and implementation in SAS[64] are also recommended.

Technical note: parameterization of treatment effects. There is a wide variety of alternative software platforms suitable for use. These range from implementations in well-known statistical packages, such as SAS, STATA, S-PLUS, or R, or variants of the WinBUGS coding suggested in this series.[2] However, the model parameterization requires care, as a number of apparently innocuous variations may give very different results or be wrong. The reviewer faced with uncited models or software devised by the investigator may need to ask for further information.[6]

C2. Multiarm Trials

C2.1 If there are multiarm trials, have the correlations between the relative treatment effects been taken into account?

When the empirical treatment differences are used as data (e.g., log odds ratios, log hazard ratios), these will be correlated in multiarm trials and the likelihood must be adjusted.[65] This is done in the Bayesian models[2] and in STATA's package mvmeta.[63] A number of software tools now under development within a frequentist framework are based on the treatment differences, and it remains to be seen whether the appropriate adjustments will be made.

C3. Connected and Disconnected Networks

C3.1. Is the network of evidence based on randomized trials connected?

It is easy to check that a network is ''connected,'' and this should be clear from a network diagram. The approach to network synthesis described in this tutorial series[2] is intended only for connected networks. Approaches used to reconnect networks require strong assumptions that must be explained and justified.[1]

C4. Inconsistency

C4.1. How many inconsistencies could there be in the network?

The network structure should be presented in a diagram[1,2,4] and the number of possible inconsistencies set out.

C4.2. Are there any a priori reasons for concern that inconsistency might exist, due to systematic clinical differences between the patients in trials comparing

treatments A and B, the patients in trials comparing treatments A and C, and so on?

If the AB trials tend to have been carried out on systematically different patient populations to the AC trials or the BC trials, there is a high risk that indirect or mixed (direct and indirect) treatment comparisons will be unreliable.

C4.3. Have adequate checks for inconsistency been made?

Different methods to check for inconsistency should be used, depending on the structure of the network.[4] A Bayesian cross-validation approach can also be used to detect the presence of outliers.[30]

C4.4 If inconsistency was detected, what adjustments were made to the analysis, and how was this justified?

If there is evidence for inconsistency in a network, it is unlikely to form a reliable basis for choosing the most effective or cost-effective treatment. A range of options are available, including removing trials from the network or incorporating additional parameters to account for bias. There are, however, likely to be a large number of ways of eliminating inconsistency, which all have quite different implications.[4]

D. Embedding the Synthesis in a Probabilistic Cost-Effectiveness Analysis

D1. Uncertainty Propagation

D1.1. Has the uncertainty in parameter estimates been propagated through the CEA model?

Failure to take account of the uncertainty in *any* parameter should be explained and justified.

D2. Correlations

D2.1 Are there correlations between parameters? If so, have the correlations been propagated through the CEA model?

Correlations between parameters are induced when they are estimated from the same data. Relative treatment effects from networks with loops are always correlated. Absolute effects of treatments based on differences from a common baseline are also correlated. Correlations must be adequately propagated through the decision model, either within Bayesian Markov chain Monte Carlo or frequentist frameworks.[6]

APPENDIX

Table A1. Checklist Table

Mark ✓ to indicate that the issue has been addressed satisfactorily and if there is any cause for concern on the item. The Comments column should be used to answer the question (YES, NO, NA: not applicable) and/or to spell out the reasons for any concerns, the need for sensitivity analyses, and so on.

		Item Satisfactory?	Comments
A. DEFINITION OF THE DECISION PROBLEM			
A1. Target Population for Decision			
A1.1	Has the target patient population for decision been clearly defined?		
A2. Comparators			
A2.1	Decision comparator set: Have all the appropriate treatments in the decision been identified?		
A2.2	Synthesis comparator set: Are there additional treatments in the synthesis comparator set that are not in the decision comparator set? If so, is this adequately justified?		
A3. Trial Inclusion/Exclusion			
A3.1	Is the search strategy technically adequate and appropriately reported?		
A3.2	Have all trials involving at least 2 of the treatments in the synthesis comparator set been included?		
A3.3	Have all trials reporting relevant outcomes been included?		
A3.4	Have additional trials been included? If so, is this adequately justified?		
A4. Treatment Definition			
A4.1	Are all the treatment options restricted to specific doses and co-treatments, or have different doses and co-treatments been ''lumped'' together? If the latter, is it adequately justified?		
A4.2	Are there any additional modeling assumptions?		
A5. Trial Outcomes and Scale of Measurement Chosen for the Synthesis			
A5.1	Where alternative outcomes are available, has the choice of outcome measure used in the synthesis been justified?		
A5.2	Have the assumptions behind the choice of scale been justified?		
A6. Patient Population: Trials with Patients outside the Target Population			
A6.1	Do some trials include patients outside the target population? If so, is this adequately justified?		
A6.2	What assumptions are made about the impact or lack of impact this may have on the relative treatment effects? Are they adequately justified?		
A6.3	Has an adjustment been made to account for these differences? If so, comment on the adequacy of the evidence presented in support of this adjustment and on the need for a sensitivity analysis.		
A7. Patient Population: Heterogeneity within the Target Population			
A7.1	Have potential modifiers of treatment effect been considered?		
A7.2	Are there apparent or potential differences between trials in their patient populations, albeit within the target population? If so, has this been adequately taken into account?		
A8. Risk of Bias			
A8.1	Is there a discussion of the biases to which these trials, or this ensemble of trials, are vulnerable?		
A8.2	If a bias risk was identified, was any adjustment made to the analysis and was this adequately justified?		

(continued)

Table A1. (continued)

		Item Satisfactory?	Comments
A9. Presentation of the Data			
A9.1	Is there a clear table or diagram showing which data have been included in the base-case analysis?		
A9.2	Is there a clear table or diagram showing which data have been excluded and why?		
B. METHODS OF ANALYSIS AND PRESENTATION OF RESULTS			
B1. Meta-Analytic Methods			
B1.1	Is the statistical model clearly described?		
B1.2	Has the software implementation been documented?		
B2. Heterogeneity in the Relative Treatment Effects			
B2.1	Have numerical estimates been provided of the degree of heterogeneity in the relative treatment effects?		
B2.2	Has a justification been given for choice of random or fixed effect models? Should sensitivity analyses be considered?		
B2.3	Has there been adequate response to heterogeneity?		
B2.4	Does the extent of unexplained variation in relative treatment effects threaten the robustness of conclusions?		
B2.5	Has the statistical heterogeneity between baseline arms been discussed?		
B3. Baseline Model for Trial Outcomes			
B3.1	Are baseline effects and relative effects estimated in the same model? If so, has this been justified?		
B3.2	Has the choice of studies to inform the baseline model been explained?		
B4. Presentation of Results of Analyses of Trial Data			
B4.1	Are the relative treatment effects (relative to a placebo or ''standard'' comparator) tabulated, alongside measures of between-study heterogeneity if an RE model is used?		
B4.2	Are the absolute effects on each treatment, as they are used in the CEA, reported?		
B5. Synthesis in Other Parts of the Natural History Model			
B5.1	Is the choice of data sources to inform the other parameters in the natural history model adequately described and justified?		
B5.2	In the natural history model, can the longer-term differences between treatments be explained by their differences on randomized trial outcomes?		
C. ISSUES SPECIFIC TO NETWORK SYNTHESIS			
C1. Adequacy of Information on Model Specification and Software Implementation			
C2. Multiarm Trials			
C2.1	If there are multiarm trials, have the correlations between the relative treatment effects been taken into account?		
C3. Connected and Disconnected Networks			
C3.1	Is the network of evidence based on randomized trials connected?		
C4. Inconsistency			
C4.1	How many inconsistencies could there be in the network?		
C4.2	Are there any a priori reasons for concern that inconsistency might exist, due to systematic clinical differences between the patients in trials comparing treatments A and B, the patients in trials comparing treatments A and C, and so on?		
C4.3	Have adequate checks for inconsistency been made?		

(continued)

Table A1. (continued)

		Item Satisfactory?	Comments
C4.4	If inconsistency was detected, what adjustments were made to the analysis, and how was this justified?		
D. EMBEDDING THE SYNTHESIS IN A PROBABILISTIC COST-EFFECTIVENESS ANALYSIS			
D1. Uncertainty Propagation			
D1.1	Has the uncertainty in parameter estimates been propagated through the CEA model?		
D2. Correlations			
D2.1	Are there correlations between parameters? If so, have the correlations been propagated through the CEA model?		

ACKNOWLEDGMENTS

We thank Phil Alderson and David Wonderling for their comments on earlier drafts of this article. We also thank Julian Higgins, Chris Hyde, Steve Palmer, Paul Tappenden, and the team at NICE, led by Zoe Garrett, for reviewing an earlier version of this paper.

REFERENCES

1. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 1: introduction. Med Decis Making. 2013;33(5): 597-606.

2. Dias S, Sutton AJ, Ades AE, Welton NJ. Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. Med Decis Making. 2013;33(5):607-617.

3. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 3: heterogeneity–subgroups, meta-regression, bias and bias-adjustment. Med Decis Making. 2013;33(5): 618-640.

4. Dias S, Welton NJ, Sutton AJ, Caldwell DM, Lu G, Ades AE. Evidence synthesis for decision making 4: inconsistency in networks of evidence based on randomized controlled trials. Med Decis Making. 2013;33(5):641-656.

5. Dias S, Welton NJ, Sutton AJ, Ades AE. Evidence synthesis for decision making 5: the baseline natural history model. Med Decis Making. 2013;33(5):657-670.

6. Dias S, Sutton AJ, Welton NJ, Ades AE. Evidence synthesis for decision making 6: embedding evidence synthesis in probabilistic cost-effectiveness analysis. Med Decis Making. 2013;33(5): 671-678.

7. Guyatt G, Oxman A, Vist G, et al; GRADE Working Group. GRADE: an emerging consensus on rating quality of evidence and strength of recommendations. BMJ. 2008;336:924–6.

8. Grades of Recommendation Assessment Development and Evaluation (GRADE) Working Group. Grading quality of evidence and strength of recommendations. BMJ. 2004;328:1490–4.

9. Briggs A, Claxton K, Sculpher M. Decision Modelling for Health Economic Evaluation. Oxford, UK: Oxford University Press; 2008.

10. Song F, Loke Y-K, Walsh T, Glenny A-M, Eastwood AJ, Altman DG. Methodological problems in the use of indirect comparisons for evaluating healthcare interventions: survey of published systematic reviews. BMJ. 2009;338(31):b1147.

11. O'Regan C, Ghement I, Eyawo O, Guyatt GH, Mills EJ. Incorporating multiple interventions in meta-analysis: an evaluation of the mixed treatment comparison with the adjusted indirect comparison. Trials. 2009;10:86. Available from: http://www.trialsjournal.com/content/10/1/86

12. Donegan S, Williamson P, Gamble C, Tudor-Smith C. Indirect comparisons: a review of reporting and methodological quality. PLoS ONE. 2011;5:e11054.

13. Higgins JPT, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 5.0.0 [updated February 2008]. Chichester, UK: The Cochrane Collaboration, Wiley; 2008.

14. Spiegelhalter DJ, Myles JP, Jones DR, Abrams KR. Bayesian methods in health technology assessment: a review. Health Technol Assess. 2000;4(38):1–130.

15. Shea BJ, Grimshaw JM, Wells GA, et al. Development of AMSTAR: a measurement tool to assess the methodological quality of systematic reviews. BMC Med Res Method. 2007;7:10.

16. Oxman AD, Guyatt GH. Guidelines for reading literature reviews. Can Med Assoc J. 1988;138:697–703.

17. Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ. 2009;339:b2535.

18. Centre for Reviews and Dissemination. Systematic Reviews: CRD's Guidance for Undertaking Reviews in Healthcare. 3rd ed. York, UK: CRD, University of York; 2009.

19. Welton NJ, Cooper NJ, Ades AE, Lu G, Sutton AJ. Mixed treatment comparison with multiple outcomes reported inconsistently across trials: evaluation of antivirals for treatment of influenza A and B. Stat Med. 2008;27:5620–39.

20. Welton NJ, Willis SR, Ades AE. Synthesis of survival and disease progression outcomes for health technology assessment of cancer therapies. Res Synthesis Methods. 2010;1:239–57.

21. Lu G, Ades AE, Sutton AJ, Cooper NJ, Briggs AH, Caldwell DM. Meta-analysis of mixed treatment comparisons at multiple follow-up times. Stat Med. 2007;26(20):3681–99.

22. Ades AE, Mavranezouli I, Dias S, Welton NJ, Whittington C, Kendall T. Network meta-analysis with competing risk outcomes. Value Health. 2010;13(8):976–83.

23. Jansen JP, Fleurence R, Devine B, et al. Interpreting indirect treatment comparisons and network meta-analysis for health-care decision making: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 1. Value Health. 2011;14:417–28.

24. Jansen JP, Fleurence R, Devine B, et al. Conducting indirect treatment comparisons and network meta-analysis studies: report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: part 2. Value Health. 2011;14:429–37.

25. Dias S, Welton NJ, Sutton AJ, Ades AE. NICE DSU Technical Support Document 1: Introduction to evidence synthesis for decision making. 2011. Available from: http://www.nicedsu.org.uk

26. National Institute for Health and Clinical Excellence (NICE). Guide to the Methods of Technology Appraisal. London, UK: NICE; 2008.

27. Hawkins N, Scott DA, Woods B. How far do you go? Efficient searching for indirect evidence. Med Decis Making. 2009;29:273–81.

28. Cooper NJ, Peters J, Lai MCW, et al. How valuable are multiple treatment comparison methods in evidence-based health-care evaluation? Value Health. 2011;14:371–80.

29. Sutton AJ, Song F, Gilbody S, Abrams KR. Modelling publication bias in meta-analysis: a review. Stat Methods Med Res. 2000; 9:421–45.

30. Dias S, Sutton AJ, Welton NJ, Ades AE. NICE DSU Technical Support Document 3: Heterogeneity: subgroups, meta-regression, bias and bias-adjustment. 2011. Available from: http://www.nicedsu.org.uk

31. Goodman S, Berry D, Wittes J. Bias and trials stopped early for benefit [letter]. JAMA. 2010;304:157.

32. Goodman SN. Systematic reviews are not biased by results from trials stopped early for benefit [letter]. J Clin Epidemiol. 2008;61:95–6.

33. Gotzsche PC. Why we need a broad perspective on meta-analysis. BMJ. 2000;321:585–6.

34. Song F, Altman D, Glenny A-M, Deeks J. Validity of indirect comparison for estimating efficacy of competing interventions: evidence from published meta-analyses. BMJ. 2003;326:472–6.

35. Song F, Xiong T, Parekh-Bhurke S, et al. Inconsistency between direct and indirect comparisons of competing interventions: meta-epidemiological study. BMJ. 2011;343:d4909.

36. Chou R, Fu R, Hoyt Huffman L, Korthuis PT. Initial highly-active antiretroviral therapy with a protease inhibitor versus non-nucleoside reverse transcriptase inhibitor: discrepancies between direct and indirect meta-analyses. Lancet. 2006;368: 1503–15.

37. Caldwell DM, Gibb DM, Ades AE. Validity of indirect comparisons in meta-analysis [letter]. Lancet. 2007;369(9558):270.

38. Caldwell DM, Ades AE, Higgins JPT. Simultaneous comparison of multiple treatments: combining direct and indirect evidence. BMJ. 2005;331:897–900.

39. Greenland S, Longnecker MP. Methods for trend estimation from summarized dose-response data, with applications to meta-analysis. Am J Epidemiol. 1992;135:1301–9.

40. Welton NJ, Caldwell DM, Adamopoulos E, Vedhara K. Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. Am J Epidemiol. 2009;169(9):1158–65.

41. Collett D. Modelling Survival Data in Medical Research. London, UK: Chapman & Hall; 1994.

42. Turner RM, Spiegelhalter DJ, Smith GCS, Thompson SG. Bias modelling in evidence synthesis. J R Stat Soc Ser A. 2009;172:21–47.

43. Dias S, Welton NJ, Ades AE. Study designs to detect sponsorship and other biases in systematic reviews. J Clin Epidemiol. 2010;63:587–8.

44. Dias S, Welton NJ, Marinho VCC, Salanti G, Higgins JPT, Ades AE. Estimation and adjustment of bias in randomised evidence by using mixed treatment comparison meta-analysis. J R Stat Soc Ser A. 2010;173(3):613–29.

45. Welton NJ, Ades AE, Carlin JB, Altman DG, Sterne JAC. Models for potentially biased evidence in meta-analysis using empirically based priors. J R Stat Soc Ser A. 2009;172(1):119–36.

46. Higgins JPT, Altman DG. Assessing risk of bias in included studies. In: Higgins J, Green S, eds. Cochrane Handbook for Systematic Reviews of Interventions Version 501 [updated September 2008]. Chichester, UK: The Cochrane Collaboration, Wiley; 2008.

47. Schulz KF, Chalmers I, Hayes RJ, Altman DG. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA. 1995;273(5):408–12.

48. Wood L, Egger M, Gluud LL, et al. Empirical evidence of bias in treatment effect estimates in controlled trials with different interventions and outcomes: meta-epidemiological study. BMJ. 2008; 336:601–5.

49. Juni P, Holenstein F, Sterne J, Bartlett C, Egger M. Direction and impact of language bias in meta-analyses of controlled trials: empirical study. Int J Epidemiol. 2002;31(1):115–23.

50. Kirkham JJ, Dwan KM, Altman D, et al. The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. BMJ. 2010;340:c365.

51. Kjaergard LL, Villumsen J, Gluud C. Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med. 2001;135:982–9.

52. Moher D, Pham B, Jones A, et al. Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet. 1998;352(9128):609–13.

53. Sutton AJ, Abrams KR, Jones DR, Sheldon TA, Song F. Methods for Meta-Analysis in Medical Research. London, UK: John Wiley; 2000.

54. Sweeting MJ, Sutton AJ, Lambert PC. What to add to nothing? Use and avoidance of continuity corrections in meta-analysis of sparse data. Stat Med. 2004;23:1351–75.

55. Bradburn MJ, Deeks JJ, Berlin JA, Localio AR. Much ado about nothing: a comparison of the performance of meta-analysis methods with rare events. Stat Med. 2007;26:53–77.

56. Higgins JPT, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. BMJ. 2003;327:557–60.

57. Spiegelhalter DJ, Abrams KR, Myles J. Bayesian Approaches to Clinical Trials and Health-Care Evaluation. New York: John Wiley; 2004.

58. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. J R Stat Soc Ser A. 2009;172: 137–59.

59. Ades AE, Sutton AJ. Multiparameter evidence synthesis in epidemiology and medical decision making: current approaches. J R Stat Soc Ser A. 2006;169(1):5–35.

60. Salanti G, Ades AE, Ioannidis JPA. Graphical methods and numerical summaries for presenting results from multiple-treatment meta-analysis: an overview and tutorial. J Clin Epidemiol. 2011;64: 163–71.

61. Cooper NJ, Sutton AJ, Lu G, Khunti K. Mixed comparison of stroke prevention treatments in individuals with nonrheumatic atrial fibrillation. Arch Intern Med. 2006;166(12): 1269–75.

62. Cipriani A, Furukawa TA, Salanti G, et al. Comparative efficacy and acceptability of 12 new generation antidepressants: a multiple-treatments meta-analysis. Lancet. 2009;373:746–58.

63. White IR. Multivariate random-effects meta-regression: updates to mvmeta. Stata J. 2011;11:255–70.

64. Jones B, Roger J, Lane PW, et al. Statistical approaches for conducting network meta-analysis in drug development. Pharm Stat. 2011;10:523–31.

65. Franchini A, Dias S, Ades AE, Jansen J, Welton N. Accounting for correlation in mixed treatment comparisons with multi-arm trials. Res Synthesis Methods. 2012;3:142–60.